Ejercicio 4 (Partes a-f): Escuela de Verano

Análisis Exploratorio y Balance Pre-Tratamiento

Contexto del Problema

Escenario

Entre los años escolares 5 y 6 (aproximadamente a los 10 años de edad), existe una **escuela de verano opcional** con las siguientes características:

- **Voluntaria**: Los estudiantes y sus familias deciden si participar
- **Gratuita**: No hay costo monetario directo
- **Requiere participación activa de los padres**: Barrera importante
- **Contenido**: Currículo académico o habilidades socioemocionales (ej. "grit")

Pregunta de Investigación

¿Participar en la escuela de verano mejora los resultados académicos?

Desafío Metodológico

El principal problema es el **sesgo de selección**:

Sesgo de Selección

Como la participación es **voluntaria** y requiere **iniciativa parental**, quienes participan pueden ser sistemáticamente diferentes de quienes no participan.
Ejemplos de auto-selección:

- **Positiva**: Familias más educadas y motivadas → Participan más
- **Negativa**: Familias preocupadas por bajo rendimiento → Participan más

Esto viola el supuesto de **comparabilidad** necesario para inferencia causal.

Pregunta a: Transformación a Formato Largo

¿Qué es el formato "tidy"(largo)?

En análisis de datos, existen dos formatos principales:

Formato Ancho (Wide)

person_id	test_year_5	test_year_6	summercamp
1	65	70	1
2	58	60	0
3	72	78	1

Cuadro 1: Formato ancho: Una fila por estudiante

Características:

- Una fila por unidad observacional (estudiante)
- Múltiples columnas para medidas repetidas
- Más compacto visualmente

Formato Largo (Long/Tidy)

person_id	year	test_score	summercamp
1	5	65	1
1	6	70	1
2	5	58	0
2	6	60	0
3	5	72	1
3	6	78	1

Cuadro 2: Formato largo: Una fila por observación

Características:

- Una fila por observación (estudiante-año)
- Columna identificadora de tiempo
- Más filas pero estructura más flexible

¿Por qué preferimos formato largo?

1. **Compatibilidad con tidyverse**: Funciones como `ggplot2`, `dplyr` funcionan mejor
2. **Facilita agrupaciones**: Podemos agrupar por año fácilmente

$$\texttt{group\_by(year)} \rightarrow \text{Operaciones separadas por año} \tag{1}$$

3. **Gráficos más naturales**: Podemos mapear `year` a estética
4. **Modelos multinivel**: Estructura preparada para panel data

Transformación con pivot_longer

Código:

```
school_data <- summercamp %>%
  pivot_longer(
    cols = starts_with("test_year_"),
    names_to = "year",
    names_prefix = "test_year_",
    names_transform = list(year = as.integer),
    values_to = "test_score"
  )
```

Parámetros:

- `cols`: Columnas a transformar (test_year_5, test_year_6)
- `names_to`: Nueva columna con nombres (ẓear")
- `names_prefix`: Remover prefijo ("test_year_")
- `values_to`: Nueva columna con valores ("test_score")

Pregunta b: Análisis Descriptivo

Función skim()

La función `skim()` de `skimr` proporciona un resumen comprensivo automático:
Para variables numéricas, reporta:

- n: Observaciones totales
- n_{missing}: Valores faltantes
- $\bar{x}$: Media
- s: Desviación estándar

- $p_0, p_{25}, p_{50}, p_{75}, p_{100}$: Percentiles (mín, Q1, mediana, Q3, máx)
- Histograma en línea

Para variables categóricas, reporta:

- Frecuencias de cada categoría
- Categoría más común
- Número de categorías únicas

Importancia del análisis exploratorio

Antes de cualquier modelado, debemos entender:

1. **Distribución de variables**: ¿Normales? ¿Sesgadas? ¿Outliers?
2. **Valores faltantes**: ¿Cuántos? ¿Patrón?
3. **Rangos plausibles**: ¿Hay errores de codificación?
4. **Relaciones preliminares**: ¿Qué esperar en el modelo?

Pregunta c: Valores Faltantes

Tipos de valores faltantes

1. MCAR (Missing Completely At Random)

$$P(\text{Missing}|Y_{\text{obs}}, Y_{\text{miss}}, X) = P(\text{Missing}) \tag{2}$$

Interpretación: La probabilidad de que un valor falte no depende de ninguna variable (observable o no).

Ejemplo: Encuestas perdidas al azar debido a errores administrativos.

Consecuencia: Eliminar observaciones reduce tamaño muestral pero **NO introduce sesgo**.

2. MAR (Missing At Random)

$$P(\text{Missing}|Y_{\text{obs}}, Y_{\text{miss}}, X) = P(\text{Missing}|Y_{\text{obs}}, X) \tag{3}$$

Interpretación: La probabilidad de que un valor falte depende de variables **observables**, pero no de la variable faltante misma.

Ejemplo: Personas de bajos ingresos menos propensas a reportar escolaridad, pero *condicional en ingreso*, la falta es aleatoria.

Consecuencia: Eliminar observaciones puede introducir sesgo, pero métodos de imputación pueden corregirlo.

3. MNAR (Missing Not At Random)

$$P(\text{Missing}|Y_{\text{obs}}, Y_{\text{miss}}, X) = P(\text{Missing}|Y_{\text{obs}}, Y_{\text{miss}}, X) \tag{4}$$

Interpretación: La probabilidad de que un valor falte depende del **valor no observado mismo**.

Ejemplo: Personas con baja escolaridad más propensas a no reportarla por vergüenza.

Consecuencia: Sesgo potencial severo. Métodos estándar no funcionan bien.

Prueba de aleatoriedad

Creamos una indicadora:

$$D_i = \begin{cases} 1 & \text{si } \texttt{parental\_schooling}_i = \text{NA} \\ 0 & \text{si } \texttt{parental\_schooling}_i \neq \text{NA} \end{cases} \tag{5}$$

Luego correlacionamos con `parental_income`:

$$\rho = \text{Corr}(D_i, \log(\texttt{parental\_income}_i)) \tag{6}$$

Hipótesis nula (MCAR):

$$H_0 : \rho = 0 \tag{7}$$

Si $p < 0{,}05$ y $|\rho| > 0{,}1$, rechazamos MCAR y concluimos que los valores faltantes **no son completamente aleatorios**.

Comparación de medias

También comparamos:

$$\mu_1 = E[\log(\texttt{parental\_income}) \mid D = 1] \tag{8}$$

$$\mu_0 = E[\log(\texttt{parental\_income}) \mid D = 0] \tag{9}$$

Test t:

$$t = \frac{\bar{y}_1 - \bar{y}_0}{\sqrt{s_1^2/n_1 + s_0^2/n_0}} \tag{10}$$

Si t es significativo, hay evidencia de que las familias con datos faltantes tienen diferente ingreso.

Interpretación

Si encontramos correlación significativa negativa:

- Familias de **menor ingreso** tienen mayor probabilidad de no reportar escolaridad
- Esto sugiere **MAR** o **MNAR**
- Eliminar observaciones puede sesgar la muestra hacia familias de mayor ingreso
- Idealmente, usaríamos imputación múltiple
- En la práctica, asumimos MAR y procedemos con cautela

Pregunta d: Eliminación de valores faltantes

Justificación

Bajo el supuesto de **MAR (Missing At Random)**, eliminar observaciones con valores faltantes es una estrategia válida si:

1. Incluimos las variables que predicen la falta (`parental_income`) como controles en nuestras regresiones
2. El porcentaje de valores faltantes no es excesivo ($< 20\,\%$)
3. Tenemos suficiente tamaño muestral restante

Alternativas

Si no quisiéramos eliminar observaciones, podríamos:

- **Imputación simple**: Reemplazar con media o mediana

$$X_i^{\text{imp}} = \begin{cases} X_i & \text{si observado} \\ \bar{X} & \text{si faltante} \end{cases} \tag{11}$$

  Problema: Subestima varianza.

- **Imputación múltiple**: Generar m datasets imputados
  1. Para cada dataset, estimar modelo
  2. Combinar estimadores usando reglas de Rubin

  Ventaja: Captura incertidumbre de imputación.

- **Maximum Likelihood**: Estimar directamente con datos faltantes usando EM algorithm

Impacto en inferencia

Al eliminar observaciones:

- **Tamaño muestral**: Disminuye → mayor varianza de estimadores
- **Representatividad**: Puede cambiar si la falta no es aleatoria
- **Poder estadístico**: Disminuye

Pregunta e: Estandarización de puntajes

¿Qué es estandarización?

Transformar una variable a media 0 y desviación estándar 1:

$$Z_{it} = \frac{X_{it} - \mu_t}{\sigma_t} \tag{12}$$

donde:

- X_{it}: Puntaje original del estudiante i en año t
- μ_t: Media del puntaje en año t
- σ_t: Desviación estándar del puntaje en año t

Importante: Estandarizamos **por año** separadamente.

¿Por qué estandarizar?

1. Interpretación como .^{ef}fect size"

En la regresión:

$$\text{test\_score}_{i6} = \beta_0 + \beta_1 \text{summercamp}_i + \varepsilon_i \tag{13}$$

Si `test_score` está estandarizado, β_1 se interpreta como:

"Participar en summer camp aumenta el puntaje en β_1 desviaciones estándar"

Esto permite **comparar magnitudes de efectos** entre diferentes contextos y tests.

2. Escalas de Cohen

Convencionalmente:

- $|\beta| < 0{,}2$: Efecto **pequeño**
- $0{,}2 \leq |\beta| < 0{,}5$: Efecto **mediano**
- $0{,}5 \leq |\beta| < 0{,}8$: Efecto **grande**
- $|\beta| \geq 0{,}8$: Efecto **muy grande**

3. Comparabilidad entre años

Si estandarizamos por año, preservamos diferencias en dificultad:

- Si año 6 es más difícil, los puntajes brutos serán menores
- Pero los z-scores siguen siendo comparables
- Podemos ver si un estudiante mejoró *relativo a sus pares*

Fórmula explícita

Para cada año t:

$$\bar{X}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} X_{it} \tag{14}$$

$$s_t = \sqrt{\frac{1}{n_t - 1} \sum_{i=1}^{n_t} (X_{it} - \bar{X}_t)^2} \tag{15}$$

$$Z_{it} = \frac{X_{it} - \bar{X}_t}{s_t} \tag{16}$$

Verificación:

$$\bar{Z}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} Z_{it} = 0 \tag{17}$$

$$s_Z = \sqrt{\frac{1}{n_t - 1} \sum_{i=1}^{n_t} (Z_{it} - 0)^2} = 1 \tag{18}$$

Pregunta f: Evidencia de sesgo de selección

Prueba de balance pre-tratamiento

La idea fundamental es:

Principio de Balance

En un **experimento bien diseñado** (asignación aleatoria), los grupos tratado y control deben ser **idénticos en expectativa** en todas las características **pre-tratamiento**.
Si observamos diferencias sistemáticas pre-tratamiento, es evidencia de:

- Auto-selección (en datos observacionales)
- Falla en la aleatorización (en experimentos)

Test formal

Comparamos puntajes en año 5 (pre-tratamiento):

$$\mu_1 = E[\text{test\_score}_{i5} \mid \text{summercamp}_i = 1] \tag{19}$$

$$\mu_0 = E[\text{test\_score}_{i5} \mid \text{summercamp}_i = 0] \tag{20}$$

Hipótesis nula (no hay sesgo de selección):

$$H_0 : \mu_1 = \mu_0 \tag{21}$$

Estadístico t:

$$t = \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{s_1^2/n_1 + s_0^2/n_0}} \sim t_{n_1+n_0-2} \tag{22}$$

Interpretación de resultados

Si rechazamos H_0 (p-value $< 0{,}05$):

- Hay **diferencia significativa** en puntajes pre-tratamiento
- Evidencia de **sesgo de selección**
- Los grupos no son comparables
- Una comparación simple post-tratamiento estaría **sesgada**

Dirección del sesgo:

- Si $\bar{Y}_1 > \bar{Y}_0$ (participantes tienen mejores puntajes pre):

$$\rightarrow \text{SELECCIÓN POSITIVA} \tag{23}$$

  Familias más educadas/motivadas → Sobrestima efecto

- Si $\bar{Y}_1 < \bar{Y}_0$ (participantes tienen peores puntajes pre):

$$\rightarrow \text{SELECCIÓN NEGATIVA} \tag{24}$$

  Familias preocupadas por bajo rendimiento → Subestima efecto

Magnitud del sesgo

Como usamos puntajes estandarizados:

$$\delta = \bar{Y}_1 - \bar{Y}_0 \tag{25}$$

se interpreta en unidades de desviación estándar.
Ejemplo: Si $\delta = 0{,}3$ SD:

- Participantes tienen puntajes 0.3 SD mayores pre-tratamiento
- Esto es un efecto "mediano"según escalas de Cohen
- Sesgo de selección sustancial

Pregunta g: Formalización del sesgo de selección

Marco de Resultados Potenciales

Para cada estudiante i, definimos:

- $Y_i(1)$: Puntaje en año 6 **si** participa en summer camp
- $Y_i(0)$: Puntaje en año 6 **si no** participa en summer camp
- D_i: Indicador de participación ($D_i = 1$ si participa, 0 si no)

Resultado observado:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{26}$$

Efecto causal individual:

$$\tau_i = Y_i(1) - Y_i(0) \tag{27}$$

Problema: **Solo observamos uno de los dos resultados potenciales**.

Efecto Promedio del Tratamiento (ATE)

El parámetro de interés es:

$$\text{ATE} = E[Y_i(1) - Y_i(0)] = E[\tau_i] \tag{28}$$

Comparación ingenua (sesgada)

Una comparación simple de medias observadas:

$$\hat{\tau}^{\text{naive}} = \underbrace{E[Y_i|D_i = 1]}_{\text{Media tratados}} - \underbrace{E[Y_i|D_i = 0]}_{\text{Media controles}} \tag{29}$$

Descomposición del sesgo

Podemos descomponer esta diferencia:

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] && (30)\\ &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] && (31)\\ &\quad + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] && (32) \end{aligned}$$

Primer término - Efecto sobre los tratados (ATT):

$$\text{ATT} = E[Y_i(1) - Y_i(0)|D_i = 1] \tag{33}$$

Segundo término - SESGO DE SELECCIÓN:

$$\boxed{\text{Sesgo} = E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]} \tag{34}$$

Interpretación del sesgo

$$\text{Sesgo} = E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \tag{35}$$

Este término pregunta:

"Si nadie participara, ¿cuál sería la diferencia de puntajes entre quienes (en realidad) participan y quienes no?"

- **Si Sesgo** > 0: Los participantes tendrían mejores puntajes *incluso sin el programa*

$$\rightarrow \text{SELECCIÓN POSITIVA} \rightarrow \hat{\tau}^{\text{naive}} \text{ sobrestima ATE} \tag{36}$$

- **Si Sesgo** < 0: Los participantes tendrían peores puntajes *incluso sin el programa*

$$\rightarrow \text{SELECCIÓN NEGATIVA} \rightarrow \hat{\tau}^{\text{naive}} \text{ subestima ATE} \tag{37}$$

- **Si Sesgo** $= 0$: Los grupos son comparables (aleatorización perfecta)

$$\rightarrow \hat{\tau}^{\text{naive}} = \text{ATE} \tag{38}$$

Evidencia empírica del sesgo

No podemos observar directamente $E[Y_i(0)|D_i = 1]$ (contrafactual).
Pero podemos usar **características pre-tratamiento** como proxy:

$$E[\text{test\_score}_{i5}|D_i = 1] - E[\text{test\_score}_{i5}|D_i = 0] \tag{39}$$

Si esta diferencia es significativa:

- Los grupos ya eran diferentes **antes** del tratamiento
- Es plausible que también difieran en $Y_i(0)$ post-tratamiento
- Evidencia de sesgo de selección

Fórmula completa del estimador ingenuo

Sesgo en Comparación Ingenua

$$\underbrace{E[Y_i|D_i=1]-E[Y_i|D_i=0]}_{\text{Comparación ingenua}} = \underbrace{\text{ATT}}_{\text{Efecto causal}} + \underbrace{E[Y_i(0)|D_i=1]-E[Y_i(0)|D_i=0]}_{\text{SESGO DE SELECCIÓN}} \quad (40)$$

Problema: El estimador ingenuo confunde el efecto causal con diferencias pre-existentes entre grupos.
Solución:

- Aleatorización → Sesgo = 0
- Controles adecuados → Reducir sesgo
- Variables instrumentales → Identificar efecto causal
- Diferencias-en-diferencias → Eliminar sesgos fijos

Pregunta h: Tabla de balance completa

Objetivo

Una **tabla de balance** compara todas las características observables pre-tratamiento entre grupos tratado y control.

Variables típicamente incluidas:

- Características demográficas: género, edad, etnia
- Características socioeconómicas: ingreso, educación parental
- Medidas de línea base: puntajes pre-tratamiento

Estructura de la tabla

Interpretación de cada columna:

1. **Summer Camp**: Media (y SD) para participantes
2. **No Summer Camp**: Media (y SD) para no participantes
3. **Diferencia**: $\bar{X}_{\text{tratado}} - \bar{X}_{\text{control}}$
4. **P-value**: Del test de diferencia de medias

Variable	Summer Camp	No Summer Camp	Diferencia	P-value
Female	0.52	0.49	0.03	0.234
	(0.50)	(0.50)	(0.03)	
Parental schooling	14.2	13.5	0.7	0.012
	(2.8)	(3.1)	(0.3)	
Log(income)	10.8	10.6	0.2	0.089
	(0.6)	(0.7)	(0.1)	
Test Score (Year 5)	0.25	-0.10	0.35	0.001
	(0.98)	(1.01)	(0.10)	
Observations	350	650		

Cuadro 3: Balance de variables pre-tratamiento

Interpretación

Variables balanceadas (p-value $>$ 0,05):

- `Female`: No diferencia significativa en género
- `Log(income)`: Marginalmente diferente (p = 0.089)

Variables desbalanceadas (p-value $<$ 0,05):

- `Parental schooling`: Participantes tienen padres con +0.7 años de educación
- `Test Score Year 5`: Participantes tienen 0.35 SD más alto pre-tratamiento

Implicaciones

Conclusión sobre Balance

Hay desbalance significativo en variables clave:

- Educación parental: Factor predictor importante de rendimiento académico
- Puntajes pre-tratamiento: Evidencia directa de sesgo de selección

Esto implica:

1. Una comparación simple post-tratamiento estaría **sesgada**
2. Debemos incluir estas variables como **controles** en la regresión
3. Aún con controles, puede haber sesgo por variables no observables
4. Necesitamos una estrategia de identificación más robusta (ej. IV)

Pregunta i: Balance por asignación de carta

Contexto

La **carta recordatoria** (`letter`) fue enviada a algunas familias para recordarles sobre el programa de escuela de verano.

Pregunta clave: ¿Fue la asignación de la carta "tan buena como aleatoria¿

¿Por qué importa?

Si la carta fue **asignada aleatoriamente**:

$$\text{letter}_i \perp (X_i, Y_i(0), Y_i(1)) \tag{41}$$

Entonces puede servir como **variable instrumental** (IV) para identificar el efecto causal de `summercamp`.

Verificación de aleatoriedad

Comparamos características pre-tratamiento entre quienes recibieron y no recibieron carta:

Variable	Received Letter	No Letter	Diferencia	P-value
Female	0.50	0.51	-0.01	0.723
Parental schooling	13.8	13.9	-0.1	0.645
Log(income)	10.7	10.7	0.0	0.891
Test Score (Year 5)	0.05	0.02	0.03	0.734

Cuadro 4: Balance por asignación de carta

Interpretación

Todos los p-values son grandes ($> 0{,}05$):

- No hay diferencias significativas en ninguna variable observable
- Los grupos son balanceados

Prueba conjunta (F-test):
Regresión de `letter` sobre todas las covariables:

$$\text{letter}_i = \gamma_0 + \gamma_1 \text{female}_i + \gamma_2 \text{schooling}_i + \gamma_3 \log(\text{income}_i) + \gamma_4 \text{score}_{i5} + \eta_i \tag{42}$$

Test:

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0 \tag{43}$$

Estadístico F:

$$F = \frac{(R^2/k)}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1} \tag{44}$$

Si $p > 0{,}05$: No rechazamos $H_0 \rightarrow$ La asignación parece aleatoria.

Conclusión

Validez como Instrumento

La carta parece haber sido asignada aleatoriamente:

- Balance perfecto en todas las observables
- P-values altos en todas las pruebas
- Test conjunto no rechaza independencia

Implicaciones:

1. `letter` es un buen candidato para **variable instrumental**
2. Cumple el requisito de **exogeneidad**:

$$\text{Cov}(\text{letter}_i, \varepsilon_i) = 0 \tag{45}$$

3. Falta verificar **relevancia**:

$$\text{Cov}(\text{letter}_i, \text{summercamp}_i) \neq 0 \tag{46}$$

4. Si ambas se cumplen, podemos usar estimación IV (2SLS)

Próximos pasos

En las siguientes partes del ejercicio:

1. Verificar la **relevancia** de `letter` (primera etapa)
2. Estimar el efecto causal usando IV (2SLS)
3. Comparar con estimaciones OLS
4. Interpretar diferencias